DOCUMENT RESUME

ED 464 950 TM 033 881

AUTHOR Shermis, Mark D.; Barrera, Felicia D.
TITLE Exit Assessments: Evaluating Writing Ability through Automated Essay Scoring.
SPONS AGENCY Fund for the Improvement of Postsecondary Education (ED), Washington, DC.
PUB DATE 2002-04-00
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002). Developed norms and project software at: http://coeweb.fiu.edu/fipsedemo.
CONTRACT P116B000387A
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Software; *Essays; Portfolios (Background Materials); *Scoring; *Test Scoring Machines; *Writing Ability
IDENTIFIERS Project Essay Grade

ABSTRACT

This paper describes ongoing work in automated essay scoring that will extend the applicability of models that are currently used for short-essay documents (i.e., less than 500 words). Sponsored by the Fund for Improvement of Post-Secondary Education (FIPSE), the project would create norms for documents that might normally be found in an electronic portfolio such as critiques, self-reflective writing, reports of empirical research, and technical reports. These norms and the software are posted on a Web site and will be made available at no cost for a period of 5 years. The paper describes the project, the desired use of electronic portfolios, and the four major automated essay scoring programs: (1) Project Essay Grade; (2) IntelliMetric; (3) Intelligent Essay Assessor; and (4) e-rater. How this technology can help evaluate postsecondary general education/principles of undergraduate learning is also discussed. (Contains 19 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 950

RUNNING HEAD: Automated Essay Scoring

Exit Assessments: Evaluating Writing Ability through Automated Essay Scoring

Mark D. Shermis Felicia D. Barrera

Florida International University

Educational and Psychological Studies

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

M. Shermis

TM033881

Paper presented at the annual meetings of the American Educational Research Association, New Orleans, LA (April, 2002)

BEST COPY AVAILABLE

Abstract

The paper describes on-going work in automated essay scoring which will extend the applicability of models that are currently used for short-essay documents (i.e., less than 500 words). Sponsored by the Fund for the Improvement of Post-Secondary Education (FIPSE), the project would create norms for documents that might normally be found in an electronic portfolio such as critiques, self-reflective writing, reports of empirical research, and technical reports. These norms and the software are posted on the website: http://coeweb.fiu.edu/fipsedemo and will be made available at no cost for a period of five years. The paper describes the project, the desired use of electronic portfolios, and the four major automated essay scoring programs. How this technology can help evaluate post-secondary general education/principles of undergraduate learning is also discussed.

Exit Assessments: Evaluating Writing Ability through Automated Essay Scoring

Introduction

This paper is designed as an update on where we are with automated essay scoring in evaluating writing at the post-secondary level (cf. Shermis & Daniels, 2001). The research and development for this grant is designed to address one small aspect of the larger problem: How do we assess undergraduate general education, or as they are sometimes called, "principles of undergraduate learning"?

For institutions that have gone through this process, faculty committees will typically identify between six and nine dimensions of general education or a similar number of undergraduate learning principles. For example, most institutions have something regarding students' ability to "reason quantitatively" or to "respect diversity". Invariably one of these principles will be: "the ability to communicate effectively". One nice thing about the principles is that there will generally be good faculty agreement as to their inclusion, akin to the consensus that one would reach by espousing democracy as a superior form of government.

However, it is in the institutional assessment of these undergraduate principles where agreement often breaks

down, since it is the details that operationalize what we mean by the construct. In the same vein, we might all espouse the virtues of "democracy", but balk at the responsibilities required by such a form of governance (e.g., the requirement to vote or tolerate extreme viewpoints).

The measurement of "communicating effectively" can take a number of forms ranging from administering an objective test to evaluating student documents written in a capstone seminar. Criticisms are typically aimed at such characteristics such as the use of restricted (departmental or unit-wide norms), insufficient or lack of information about validity and reliability, reliance on idiosyncratic rubrics, and failure to identify factors contributing to student growth in progressing throughout the program.

A measurement procedure that holds some promise in overcoming these difficulties is the electronic portfolio. Similar to typical portfolios, it is a purposeful organization of learner-selected evidence of school and non-school accomplishments, but stored on electronic media including floppy disks, CD-ROMs, or the World Wide Web. The definition has several important components. First, the phrase "purposeful organization" suggests that the

"evidence" contained in the portfolio constitutes something more than a "grab bag" of materials. Usually the work represents the best example of what the learner is capable of doing for a particular class of products. For example, a psychology major might place a report of an empirically-based experiment in her portfolio as exemplary work for an undergraduate. It would not be unreasonable for faculty to suggest what classes of products would generate compelling evidence of good or excellent work. Moreover, in order to employ portfolios (or any assessment technique for that matter), faculty need to have established and communicated learning objectives developed at the departmental level.

The second important component of the definition suggests that the selections are made by the student. This means that sometime during their education, students would have to develop criteria and expertise to evaluate their own work. In this light, Stemmer (1993) relates five of the six major premises underlying the use of portfolios to include: (1) Is learner-centered and learner-directed; (2) Is a developmental tool to help the learner set goals and expectations for performance; and (3) Is an instrument that provides a means for the learner to become self-aware and capable of gathering stronger evidence of skills (4); Is a basis for documenting and planning lifelong learning; and

(5) constitutes an integration of career planning, counseling, curriculum, instruction and assessment activity.

Finally, the definition of portfolios suggests that selections might come from outside the formal curriculum. For example, a psychology major might list volunteer work from a HeadStart program as part of her portfolio. This work would not only be relevant with regard to the values inculcated by the institution for the purpose of service learning, but the choice itself would be related to the major. Stemmer (1993) reiterates this when he states that the sixth premise of using electronic portfolios is (6) to be inclusive of the entire program.

Portfolios have become increasingly popular over the past few years. They have been in use in disciplines where portfolios have a long-standing presence (e.g., marketing, communications, graphic arts). They are also becoming very popular in disciplines that are aligned with professional schools (e.g., education or business). The use of electronic portfolios has emerged as a popular alternative in smaller liberal arts colleges (e.g., Kalamazoo College, Winona State University).

In addition to addressing the lack of consensus on measurement formats, portfolios have a number of

advantages. First, they typically engender greater "buy-in" from both students and faculty. Second, they have secondary utility for other purposes such as job interviews or applications for graduate school. Third, portfolios are applicable for both individual and program evaluations. Finally, portfolios have historically had good correlations with outcome measures in disciplines where they have been used (Hatfield, 1997).

Disadvantages include: (1) significant time investments for faculty and students, especially during startup activities; (2) lowest common denominator expectations regarding hardware and software sophistication of faculty and students; (3) requirements for technical support; and (4) varying acceptance levels from other potential consumers.

Portfolios got their start in smaller, liberal arts institutions where the faculty-to-student ratio is low, writing is a significant component of the undergraduate curriculum, and there is strong faculty commitment to engage in a process of continuous program improvement. Moreover, assessing portfolios is often a labor-intensive operation. Faculty are required to evaluate a variety of documents using either a host of different rubrics. If a

department has 500 majors, using a portfolio can be a daunting challenge.

Institutions that employ portfolios have to decide whether to sample from a pool of available portfolios or to evaluate everyone's portfolio. If sampling is used, then it is often hard to establish student compliance since at least a few will believe that there is no way that they could get chosen. If everyone is evaluated (e.g., because of an accreditation or licensing requirement), then sheer numbers can make the task difficult.

One mechanism that might be used to address the labor issue of grading portfolios, especially in electronic form, is automated essay scoring—a relatively recent technological development. It holds promise for establishing national norms against which writing performance might be evaluated, formulating developmental norms that would allow an institution to track changes in student writing quality over time, and incorporating a mechanism for using formative feedback in literacy (writing) instruction (Shermis & Daniels, in press).

Automated Essay Scoring: What is it?

Automated essay scoring (AES) engines employ computer technology to evaluate and score written prose. Although most research on this technique has involved the English

language, models are being developed concurrently for evaluation of other languages (Shermis & Burstein, in press). Not all writing genre are included in this definition, and indeed, we suspect that certain ones may never be covered (e.g., poetry). Nonetheless, it is estimated that approximately 90% of required writing in a typical college classroom can be evaluated through AES.

In AES grading, rater behavior is used as the ultimate criterion, though at least one system (Intelligent Essay Assessor- Landauer, Laham & Foltz, in press) evaluates content on the basis of external material. Bennett and Bejar (1998) in criticizing the over-reliance on human ratings as the sole criterion for evaluating computerized assessment performance, claim that such ratings, typically based on a within domain constructed rubric, that may ultimately achieve acceptable reliability, but at the cost of external validity. They suggest that three issues must first be addressed in order to maximize the validity of the rating process: First, there is no theory per se for what constitutes good writing, so using an evaluation scheme in a vein suggested by Messick (1989) is difficult. Second, it appears as if "good writing" rules are made to be broken. It is only when the writer violates general rules of grammar and syntax that a consensus can be formulated

concluding that the writing is less than satisfactory. In this light, even with substantial training and good evaluation rubrics, high reliability of ratings among humans is hard to achieve. Third, even when good reliability among human raters is obtained, it is sometimes for different reasons. The best conclusion that can be reached is that it is hard to get raters to articulate why an essay is good (or bad), but that they can recognize good writing when they see it (Shermis, Koch, Page, Keith, & Harrington, 2002).

Page and Peterson (1995) discuss the use of proxes and trins as a way to think about the process of emulating rater behavior. Trins represent the characteristic dimension of interest such as fluency or grammar whereas proxes (taken from approximations) are the observed variables with which the computer works. These are the variables into which a computer parser might classify text (e.g., part of grammar, word length, word meaning, etc.). In social science research, a similar distinction might be made between the use of latent and observed variables.

In terms of its present development, one might think of AES as representing the juncture between cognitive psychology and artificial intelligence. The AES engines, described in the following section, demonstrate that the

correlation of technology with human rater behavior. The AES engines, predict as well or better than scores produced by raters, and yields a high degree of construct validity. Explanations as to why it works well are only beginning to emerge as implicit or tacit "trade secrets", and may not correspond well to past research (Shermis & Burstein, in press). Accordingly, the technology must be viewed "in the making" akin to where microcomputers were in the early 1980's, impressive for the time being, but having the potential for improvement.

The AES Scoring Engines

Project Essay Grade:

The first automated essay scorer to be developed was Project Essay Grade (PEG; Page, 1966). Although initial work on PEG started in the 1960's, some practical problems weren't solved until the microcomputer became popular in the late 1980's, at which time ETS conducted a blind test of PEG for scoring 1,314 essays produced by students taking the Praxis test, used in evaluating applicants for teacher certification (Page & Peterson, 1995). The results supported the hypothesis that PEG was more accurate in predicting human ratings up to and including three human judges (Page & Peterson, 1995). In essence, the automated

grading of essays proved to be not only more accurate, but also more rapid and economical. By the same token, past work on PEG has yielded favorable results when studying the traits within an essay (e.g. it's style, content, and creativity). One recommended use of such traits according to Page (in press) would be "to apply them ipsatively, i.e., comparing the traits as measured within the student". This type of evaluation would yield information as to what trait a specific student is especially strong in and which they need to improve; proving to be an invaluable tool for the improvement of writing skills.

Since the early 1990's, PEG technology has been modified in several ways. Since then it has acquired several parsers and dictionaries and it has incorporated special collections/classification schemes (Page, in press). Shermis, Mzumara, Olson, & Harrington (2001) reported on PEG's use of a web-based interface for grading student placement test essays. The design consisted of 1200 essays scored holistically by four different raters. The results were encouraging; human judges correlated .62 percent of the time, while PEG correlated with the judges at .71. In addition, the grading speed of PEG improved to grading about three essays per second (Shermis et al., 2002). In sum, PEG has resulted in a very efficient and

economical project that has radically improved the functionality of automated essay grading.

Intellimetric

IntelliMetric, a second type of automated essay scorer, has also been shown to be highly effective. It was first made available to educational agencies in January of 1998 and was the first essay-scoring tool based on artificial intelligence.

Intellimetric relies on Vantage Learning's CogniSearch™ and Quantum Reasoning™ technologies, the specific characteristics associated with each score point are internalized and then applied to subsequent scoring. Interestingly, the scoring engine may be said to "learn" which characteristics raters tend to value highly and those that the raters associate with poor scores.

IntelliMetric technology parallels processes of holistic scoring and human raters: e.g. on the one hand, human scorers trained to be prompt-specific, and, on the other, Intellimetric is able to create a solution for each stimulus prompt (Elliott, in press). It is capable of analyzing English into seventy-seven semantic, syntactic, and discourse level features (Elliott in press) in five different categories: focus and unity, development and elaboration, organization and structure, sentence

structure, mechanics and conventions. These are to be extended to other languages including French, Dutch, Portuguese, and Italian some time in 2002.

IntelliMetric is based on the merging of artificial intelligence, natural language processing, and statistical technologies. According to Elliott (in press), "It internalizes the pooled wisdom of many expert scorers." It has been used to score open-ended, essay-type questions in English, Spanish, Hebrew and Bahasa (Elliott, in press).

IntelliMetric uses a multi-stage procedure to score essay-type responses. In the first step, IntelliMetric internalizes the known score points of a set of responses. Subsequently, the model is tested against a smaller set of responses with known scores that aides in validation and generalizability of the model. Once these are confirmed, the model is used to score new responses whose scores are unknown. Responses are targeted if they are evaluated to be atypical with regards to the standards previously set by the essay scoring or by standard American English.

IntelliMetric may be applied in either "Instructional" or "Standardized Assessment" modes. As an Instructional tool, it provides feedback on a specific student's overall performance. In particular, it provides diagnostic feedback on several dimensions like organization and on analytical

dimensions like sentence structure (Elliott in press). It permits a student to revise and edit their own essay compositions. The Standardized assessment mode, is configured to provide for a single student's submission with a holistic score and if need be, feedback on several rhetorical and analytical dimensions (Elliott, in press).

With regards to the validity of IntelliMetric, various designs have been employed that fall within three main categories. One is the IntelliMetric-Expert Comparison Studies, which provides comparisons between IntelliMetric's scores and those produced by about two expert raters. The second is the True Score Studies which uses a large number of expert raters, whose scores are then averaged and used as a proxy for the true score. This true score approximation is then compared to the IntelliMetric score and the experts' scores. The third category is that of Construct Validity Studies, in which the both the scores produced by IntelliMetric and expert raters are compared to other external measures to evaluate whether IntelliMetric is consistent with the expectations for the construct (Elliott, in press). In sum, IntelliMetric has showed greater accuracy in scoring than expert raters (Eliott, in press).

Intelligent Essay Assessor

The third essay scoring system in the development of AES is that of the Intelligent Essay Assessor (IEA). Based on Latent Semantic Analysis (LSA), it is used for scoring the quality of both conceptual content-based essays and creative narratives. Most importantly, LSA technology provides direct, content-based feedback to instructors or teachers (Landauer, Laham & Foltz, 1998). "LSA provides a representation of an essay's semantic content as a vector (e.g. a set of factor loadings) computed from a set of words contained in the essay. Each vector is compared with another through a cosine, for comparing similarities (Landauer et al., in press). The vector length is defined as the length of each point from the origin.

LSA technology uses three different methods for evaluating both the quality and quantity of knowledge within an essay. They are 1) pre-scored essays of other students; 2) expert model essays and knowledge source materials; 3) internal comparison of an un-scored set of essays (Landauer et al., in press). These methods provide information regarding the degree to which a specific student's essay has content of the same meaning as that of the comparison texts).

The primary method of evaluation, Holistic, compares an essay of unknown quality to a set of pre-scored essays. "In LSA, vectors are used to produce two independent scores, one for the semantic quality of the content, the other for the amount of such content expressed" (Landauer et al., in press,). A quality score is derived by having human raters score a large sample of student essays. Subsequently, each of the human-scored essays is compared with the to-be-scored essays. Then about ten of the pre-scored essays that most resemble the specific target essay are selected. Finally, this target essay is given "the weighted-by-cosine-average human score of those in the similar set" (Landauer et al., in press).

In particular, the Intelligent Essay Assessor has proven to be very useful for not only quick and efficient essay scoring, but also for detecting plagiarism. Since every essay is compared to every other essay in a given set, if two are found to be similar they are flagged by IEA (Landauer et al., in press). This may prove to be an invaluable tool for educators that do not have the ability, with 150 or more essays to grade, to detect students' plagiarism. Since this form of academic dishonesty is so hard to detect by human scorers, automated essay scoring

technology may shed light into a previously illusive concept.

In sum, IEA's future consists in expanding beyond the more global assessment of such characteristics like flow and coherence to more specific ones like audience focus and voice (Landauer et al., in press). Consequently, these improvements may result in the expansion of IEA technology for assessment purposes.

E-Rater

The final essay scoring system is *e-rater*, developed by the Educational Testing Service (ETS) in 1999 for the operational scoring of the GMAT Analytical Writing Assessment. In use, examinees are assigned an *e-rater* score and one human reader score, a process used to score about 360,000 essays per year (Burstein, in press). The reported discrepancy rate on these massive sets of data has been less than three percent (Burstein, in press), demonstrating that *e-rater* technology is a reliable measure of essay scores.

The *e-rater* scoring system aims to implement similar features to those used in holistic scoring. It's scoring is based on a six-point scale. To score on the higher end of the scale, an essay must remain consistent with its topic and have a strong, well-organized argument. In addition, an

essay must also consist of a strong syntactic structure and use a diversity of words (Burstein, in press). "*E-rater* features include discourse structure, syntactic structure, and analysis of vocabulary usage (topical analysis)", [but do] "not include direct measures of length, such as word count in essays, or transformations of word count" (Burstein, in press).

Recently, *E-rater* has been incorporated with *Criterion*^{SM}, which is an online, web-based, essay evaluation project of ETS Technologies, a for-profit subsidiary of ETS. This project is currently used by "institutions for high- and low-stakes writing assessment, as well as for classroom instruction" (Burstein, in press). As a result, teachers and students are able to view the *e-rater* score and feedback within seconds.

In sum, *e-rater* scores essays based on a prompt-specific model (Burstein, in press). More recent research focuses on the development of more generic, global *e-rater* scoring models. Furthermore, work is being pursued to provide meaningful scores to specific essay traits, like organization, content, and style, as opposed to overall holistic scores (Burstein, in press). Current research in automated essay scoring has indicated that *e-rater* performs comparably to human readers at different grade levels

(Burstein, in press). She also reported that *E-rater* models exist for prompts based on data samples from grades 4 through 12 using national standards prompts; for undergraduates, using English Proficiency Test (EPT) and PRAXIS prompts; and, for non-native English speakers, using TOEFL prompts. ETS programs, including GMAT, TOEFL, and GRE are currently using *e-rater* with *Criterion* for low-stakes, practice tests (Burstein, in press).

The FIPSE Project

Shermis (2000) has designed a FIPSE-funded project to create national norms for documents found commonly in electronic portfolios. These norms will then be available, for a period of five years, through automated software that could grade the documents via the World Wide Web. Documents to be included in the norming procedure have been drawn from four writing genre: reports of empirical research, technical reports, historical narratives, and works of fiction.

This application is based on previous research with shorter (i.e., less than 500 words) essays in which computers have surpassed both the reliability and validity of human raters. The ultimate criterion in this process are the evaluations of human raters, and the results of regression models of writing based on large numbers of

essays and raters. In order to build the statistical models to evaluate the writing, several institutions from across the country, representing a range of Carnegie classifications, have agreed to provide 400-750 documents that are reflective of their current electronic portfolios. Six raters will evaluate each document and provide both holistic and trait ratings.

Vantage Technologies, Inc. has agreed to provide their Intellimetric™ parser for both model building and actual implementation of the project. Post-secondary institutions that are moving towards electronic portfolios could benefit from having access to the comparative information. Moreover, establishing norms would allow an institution to examine writing development of students over time. Finally, the software could be used in a formative manner, allowing students to preview their writing evaluations in order to improve writing or make better document selections.

Because previous work with the Intellimetric™ grading engine placed a heavy emphasis on content, and needed to be modified to focus on the characteristics of general writing ability, we ran a study to determine to which it would score as reliably as other engines (Shermis et al., 2002). We also needed to test the ability of the Intellimetric™

engine to interact with our web-based support mechanisms. The study was conducted in October and finalized in November of 2000 (Vantage Learning, 2000). The results showed that the modifications to the Intellimetric™ engine resulted in inter-rater agreement coefficients that were as high, and in a few cases, higher than the AES software we had worked with in the past. Moreover, the web-based support mechanism that we had used for previous work was easily adaptable to the Intellimetric™ engine. So that prospective users might give the software a "tryout, we have set up a demonstration web site based on a placement testing application (the model assumes a document of 500 words or less)s which is located at:

http://coeweb.fiu.edu/fipsedemo

This website is publicly available and can be used for the grading of shorter essays, though this aspect of the web site is not actually part of the proposed project.

Over the past year, we have been collecting data on the four genres mentioned above. Although most of the participating institutions are moving towards electronic processing of documents, sites contributing to our normative database often supply paper copies of documents. Even though we are grateful for the data, the papers are labor-intensive to handle since they require scanning,

optical-character recognition processing, editing, and coding. We are positioned to begin evaluation ratings for the genres of critiques and self-reflective writing, but are looking for additional documents in the areas of technical- and empirical report-writing. In addition to providing holistic scores, six raters will be recording trait ratings using Northwest Educational Research Laboratory's 6+1 Traits™ rubric (NWREL, 1999). For those who are not familiar with AES, we two resources have been developed for understanding the new technology. The first is a website:

http://coeweb.fiu.edu/webassessment/aes

which discusses some of the concepts involved in AES, links to many of the authors (and their work) who are doing AES research, links to the AES grading engine demonstration sites, and provides a bibliography of work in the area. The second resource is a forthcoming edited book in automated essay scoring entitled, Automated Essay Scoring: A Cross-Disciplinary Perspective published by Lawrence Erlbaum Associates, Inc. . It is designed as a primer on the topic and an exploration into future possibilities for AES.

Conclusions

In this paper, we have provided background information on what automated essay scoring is, a brief review of four popular automated essay scoring engines, and an update to a FIPSE-sponsored project that incorporates automated essay scoring into electronic portfolios.

If this project is successful, then it may simply be a matter of some minor programming to incorporate the AES models described herein as part of a distance learning package (for formative use) or as component of an institutional portfolio that monitors student progress on principles of undergraduate learning (a summative use).

Employing national norms for automated essay grading in this fashion can supplement locally-developed human-administered rubrics that focus on content in the major or indicators for program improvement. AES, as described here, is not meant to preclude assessment by humans, but makes possible a more thorough evaluation of students' written work. This information can be very helpful for improving writing, modifying programs of instruction, or making some global assessment of the state of general education in an institution.

We welcome inquiries from other interested individuals or institutions regarding this project and would be happy to work with you in applying AES in your assessment plans.

Author Notes

Correspondence concerning this article should be addressed to Mark D. Shermis, Florida International University, ZEB 310 University Park, Miami, FL 33199. Electronic mail may be sent via Internet to MShermis@FIU.Edu. Research for this project was sponsored by the Fund for the Improvement of Post-Secondary Education (FIPSE Grant # P116B000387A). The opinions expressed in this paper do not necessarily reflect those of FIPSE or the U.S. Department of Education.

References

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Burstein, J. (in press). The E-rater™ Scoring Engine: Automated Essay Scoring With Natural Language Processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach* Mahwah, NJ: Lawrence Erlbaum.

Elliot, S. (in press). Intellimetric™: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach*. Mahwah, NJ: Lawrence Erlbaum.

Hatfield, S. (1997, November). *Assessment in the major: Tools and tips for getting started*. Paper presented at the Assessment Conference in Indianapolis, Indianapolis, IN.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998) An introduction to latent semantic analysis. Discourse Processes, 25, 2&3, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (in press). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach.* Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: MacMillan.

NWREL. (1999, December). *6+1 Traits™ of Writing Rubic* [web site]. Northwest Educational Research Laboratory. Retrieved, from the World Wide Web: http://www.nwrel.org/eval/pdfs/6plus1traits.pdf

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan,* 47, 238-243.

Page, E. B. (in press). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach.* Mahwah, NJ: Lawrence Erlbaum.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(6), 561-566.

Shermis, M. D. (2000). *Automated essay grading for electronic portfolios* (Grant No. P116B000387A).

Washington, DC: Fund for the Improvement of Post-Secondary Education.

Shermis, M. D., & Burstein, J. (in press). *Automated essay scoring: A cross disciplinary approach.* Mahwah, NJ: Lawrence Erlbaum.

Shermis, M. D., & Daniels. (2001) Automated essay grading for electronic portfolios. *Assessment Update,* 13 (1), 10.

Shermis, M. D., & Daniels. (in press) Web applications in assessment. In T. W. Banta (Ed.). *Building a Scholarship of Assessment*. San Francisco: Jossey-Bass.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62* (1), 5-18.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education, 26*(3), 247-259.

Stemmer, P. (1993, February). *Electronic portfolios: Are going to the very next craze.* Paper presented at the Michigan School Testing Conference, Ann Arbor, MI.

Vantage Learning (2000). A true score study of Intellimentric™ accuracy for holistic and dimensional

scoring of college entry level writing responses.

Author: Yardley, PA. (RB-407).